Tutorial

Variant Detection

April 10, 2026

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Sample to Insight

Tutorial

Variant Detection

Introduction

The purpose of this tutorial is to demonstrate how *CLC Genomics Workbench* and the Biomedical Genomics Analysis plugin can be used to detect somatic variants at low frequencies in targeted sequencing data.

We focus on the following:

- Import the data.
- Detect variants using a template workflow.
- Interpret the results.
- Optimize variant detection.
- Optionally upload results to QCI Interpret.

Data used in this tutorial

This tutorial uses a sample containing the Seraseq® Myeloid Mutation DNA Mix reference material, which contains 23 verified clinically relevant variants, spiked into the HG002 genome in a bottle background. The sample was prepared using a QIAseq Multimodal Panel.

To complete the tutorial in a reasonable amount of time, only a subset of reads mapping to chromosomes 13 and 19 are used here, which contains six of the 23 verified variants (table 1).

Chr	Gene	Type	Nucleotide change	Protein change	VAF
13	FLT3	Insertion	c.1759_1800dup	p.N587_D600dup	5%
13	FLT3	Insertion	c.1806_1807insGG...AA	p.K602_W603insGAFREYEYDLK	10%
13	FLT3	SNV	c.2503G>T	p.D835Y	10%
19	CALR	Deletion	c.1099_1150del	p.L367Tfs*46	5%
19	CEBPA	Insertion	c.68dup	p.H24Afs*84	15%
19	CEBPA	Insertion	c.937_939dup	p.K313dup	15%

Table 1: The six variants in the reference material located on chromosomes 13 and 19.

The data distributed for use with this tutorial contains:

- The reads from the sample described above.
- A variant track containing the six expected variants (table 1).
- A variant track containing the background variants from HG002.
- An annotation track containing the high confidence regions of HG002.
- A workflow for assessing the recall and precision of the detected variants.

Tutorial

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench 26.0.1* and Biomedical Genomics Analysis *26.0.1* or higher. Note that higher versions may produce slightly different results than those shown here.

Installing plugins is described in the *CLC Genomics Workbench* manual.

General tips

- Throughout this tutorial, we provide links to relevant manual pages, which we recommend exploring for additional details.
- Tools and workflows can be found in the Toolbox, but it is often easier to launch them using Quick Launch (), found in the top toolbar (shortcut Ctrl+Shift+T or ⌘ +Shift+T on Mac). Quick Launch displays the full Toolbox path, making it easy to identify the location of the tool or workflow if needed.
- The in-built manual can be accessed by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.
- Within wizards, the **Reset** button can be used to change settings to their default values.
- Columns in tables can be hidden by unchecking their name in the Side Panel.
- Columns in tables can be used to sort the rows, by successively clicking on the column name until the desired order (indicated by an arrow next to the column name) is achieved.
- Many data elements produced by *CLC Genomics Workbench* tools have multiple views, indicated as icons in the lower left corner of elements opened in the View Area. Clicking on one of the view icons while pressing the Ctrl (⌘ on Mac) key will open in split view such that both views are visible at the same time. Often, if viewing a table and a graphical representation in split view, selecting entries in the table will highlight them in the graphical representation. The order of the views can be changed using drag and drop, see Arrange views in View Area.

Import the data

We start by downloading and importing the tutorial data:

1. Download the tutorial data.
2. Start *CLC Genomics Workbench*.
3. Import the tutorial data using Standard Import:
   (a) Launch **Standard Import** ( ) using Quick Launch ( ).
   (b) Locate the data using the **Add files** button and select **Automatic import** (figure 1).

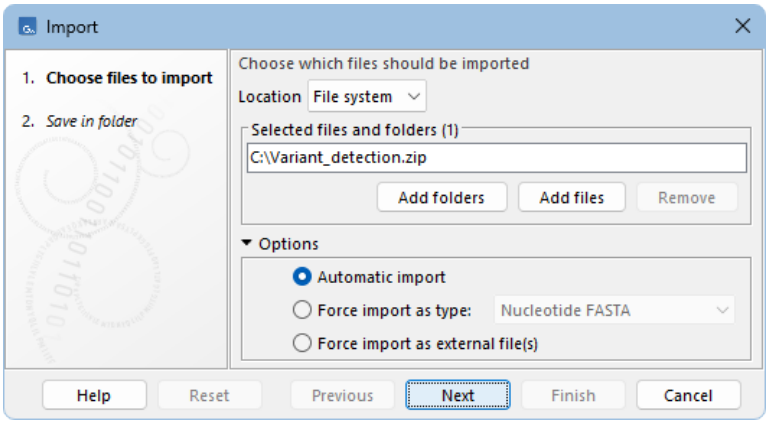

Figure 1: *Standard Import configured to import the tutorial data.*

   (c) In the next step, select a suitable location in the Navigation Area to save the imported data and click on **Finish**.

Once the import is completed, the tutorial data is visible in the Navigation Area (figure 2).

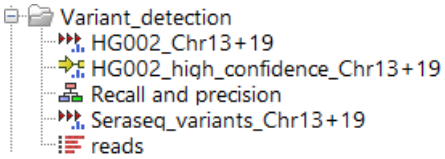

Figure 2: *The imported tutorial data in the Navigation Area.*

Detect variants

We will now use the Analyze QIAseq Multimodal DNA Panel template workflow to analyze the tutorial data. This workflow has been designed for data generated using a QIAseq Multimodal Panel. The Biomedical Genomics Analysis plugin provides several template workflows for variant detection, covering a range of data types and applications. If you run any template workflow on your own data, please note that they are provided as example workflows and may need to be customized to meet the specific requirements of your data.

To see the content of the workflow, locate it in the Toolbox:

Workflows | **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq DNA workflows** ( ) | **Analyze QIAseq Multimodal DNA Panel (Illumina)** ( )

right-click on its name and choose **Open Copy of Workflow**.

We will now run the workflow:

1. Launch the workflow using Quick Launch ( ) or by double-clicking its name in the Toolbox.
2. In the first wizard step, **Specify workflow path**, keep the default settings.
3. In the **Select Reads** step, select the imported reads (figure 3).

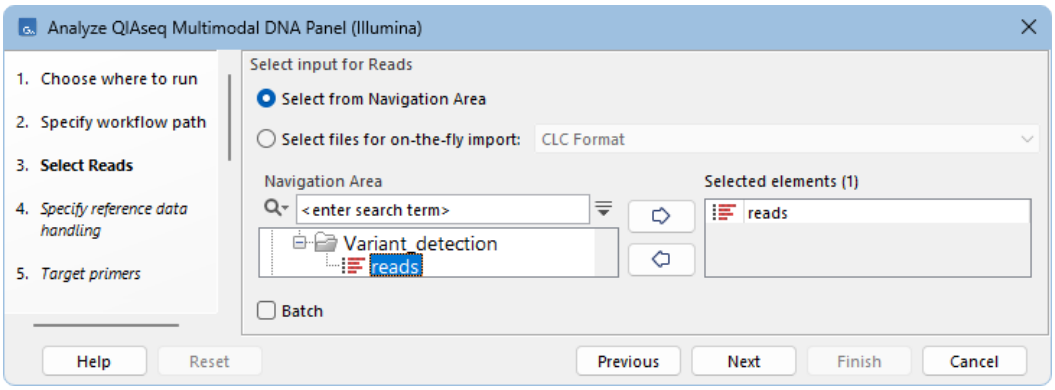

Figure 3: *The imported reads are used as input.*

4. In the **Specify reference data handling** step, select the "Variant Detection" Reference Data Set under "QIAGEN Tutorial" (figure 4). Click on **Download to Workbench** if the set has not already been downloaded.

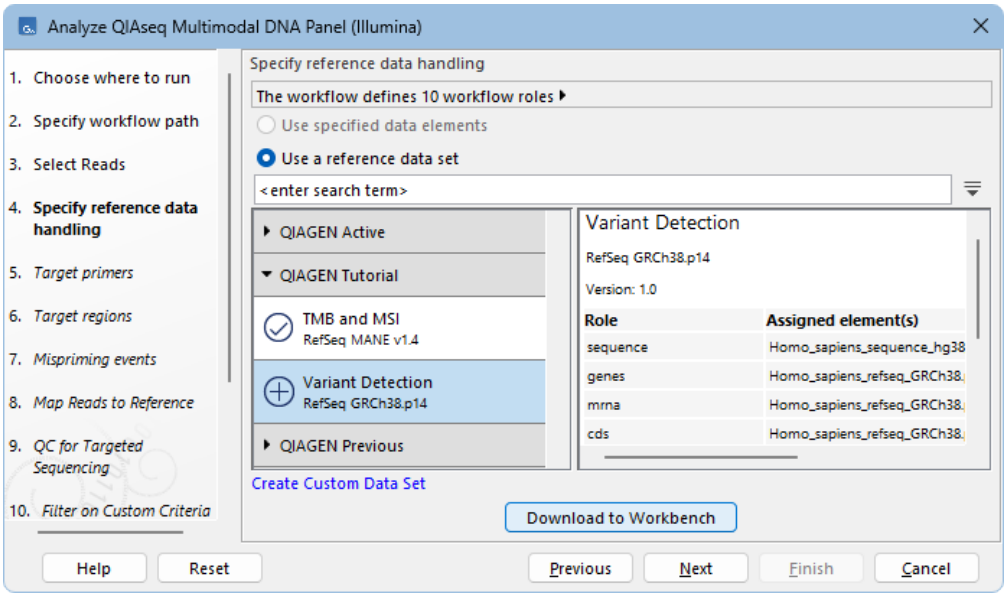

Figure 4: *The "Variant Detection" tutorial reference data set is selected and ready to be downloaded.*

5. In the next steps, keep the default settings.
6. In the **Filter on Custom Criteria** step, expand the "Low quality or support" group (and its subgroup) and set **Frequency** to 2.5 (figure 5), as this value is half of the minimum frequency of the variants in the reference material (table 1). Before clicking on **Next**, save

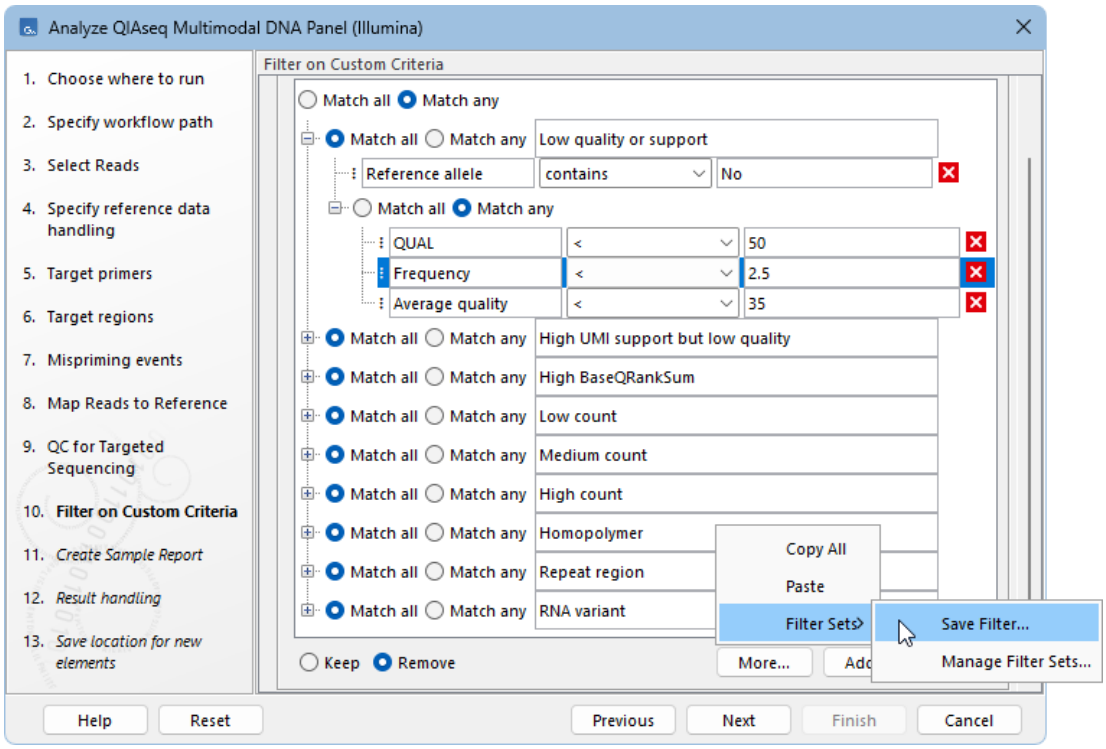

Figure 5: *The frequency is set to 2.5 and the filter criteria are ready to be saved as a filter set.*

the filter criteria to a set by clicking on **More...**, **Filter Sets**, and **Save Filter...**, and name the set "Filter variants".

7. In the last step, make a new subfolder in "Variant_detection" called "Results" and choose to save the workflow results there.

   Click on **Finish**.

The workflow will now execute. The progress can be monitored under the Processes tab in the Toolbox.

Interpret the results

Results from the workflow are placed in the "Results" folder (figure 6).

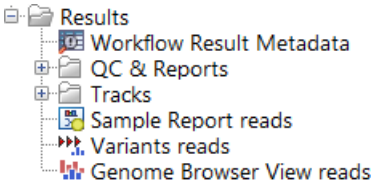

Figure 6: *The "Results" folder in the Navigation Area.*

The folder contains, among other things:

- A **QC & Reports** subfolder containing all reports produced by the workflow.
- A **Tracks** subfolder containing the tracks produced by the workflow, including the reads track.

- A Variant track ( ) containing the detected variants.
- A Genome Browser View ( ) containing various tracks produced by the workflow, including the reads track and detected variants.
- A Sample Report ( ) summarizing information from many of the reports located in the "QC & Reports" subfolder.

Quality control

It is important to first verify that the data quality is satisfactory before investigating the detected variants.

Open the **Sample Report reads**, located in the "Results" folder. The Quality control section contains different summary items that can be used to assess the quality of the reads (figure 7). These summary items can be configured in the "Create Sample Report" wizard step when launching the workflow.

1.2 Quality control

Summary item	Report type	Value	Threshold
Percentage reads mapped in target region	Raw read coverage	81.65	≥ 50.00
Percentage of target region positions with coverage ≥ threshold	UMI read coverage	0.44	≥ 90.00

Figure 7: *The sample report contains summary items for assessing the quality of the reads.*

As the tutorial data is a subsample of the full sequencing reads, a low percentage of targets have sufficient coverage, and the "Percentage of target region positions with coverage ≥ threshold" summary item is marked in yellow. This is expected for the tutorial data and does not affect the analysis performed here.

The subsequent sections provide more details about the sample analysis. For example:

- The reads have good quality scores, as few reads were trimmed for low quality (section: *Trim adapters* | *Detailed trim results* | *Trim on quality*).
- A high percentage of reads mapped to the reference genome (section: *Map reads to reference* | *Reads summary*).
- There is on average just one read per UMI group (section: *Calculate unique molecular index groups* | *Groups*).
- The median UMI coverage is just above 160 (section: *UMI read coverage* | *Summary*).

See Quality control of DNA reads with UMIs for more information.

Detected variants

Open the **Genome Browser View reads**, located in the "Results" folder. This automatically opens a split view, with the browser at the top and the table view of variants at the bottom (figure 8).

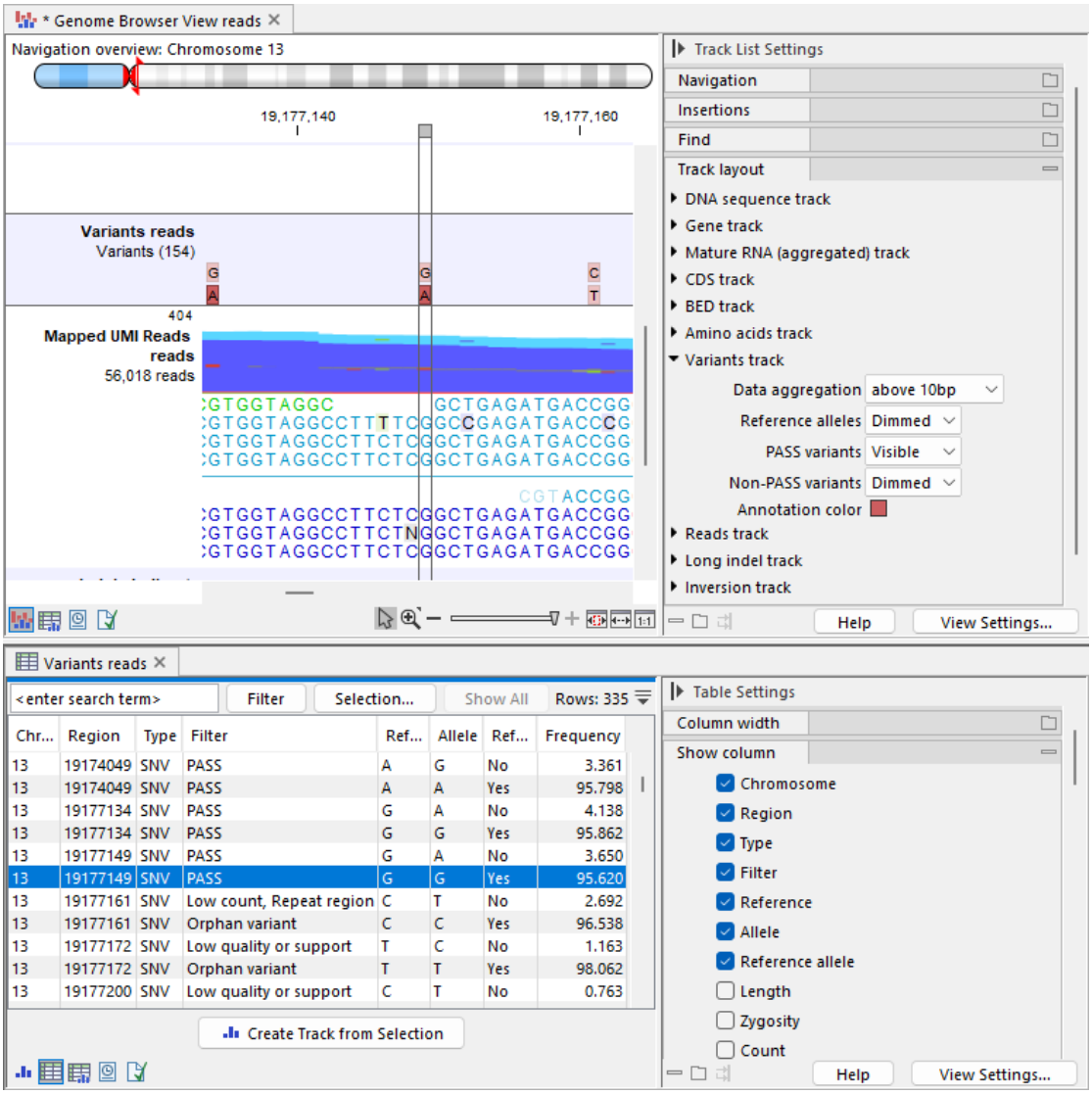

Figure 8: *Genome Browser View opened in split view. Top: The tracks order in the browser has been changed by dragging the reads track immediately below the variant track. The reference alleles and non-PASS variants are dimmed from the Side Panel. Bottom: The table view of variants. The "Filter" column indicates if a variant passed all filters, or which filters it did not pass. A variant is selected, causing the browser to zoom in to that variant. The table view does not show all columns.*

The Genome Browser View produced by template workflows typically contains gene, mRNA, CDS, and reads tracks, along with the variant track and the resulting amino acid changes. Clicking on a row in the table view automatically zooms the browser to the corresponding location, providing the full genomic context in that region. Hovering the mouse over any variant in the browser reveals a tooltip with detailed information.

Both the table ( ) and track ( ) views of the variant track ( ) show all variants by default. This includes reference variants and variants that did not pass the filters configured in the "Filter on Custom Criteria" wizard step when launching the workflow. The information on whether the variant is a reference variant is located in the "Reference allele" table column, while the "Filter" column contains:

- "PASS" if the variant passed all filters.
- A description of the filters the variant did not pass.
- "Orphan variant" if the variant is a reference variant and the non-reference variant(s) at the same position did not pass the filters.

To focus on the non-reference alleles that have passed all filters:

- In the "Track layout" palette of the browser's Side Panel, expand the "Variants track" group, and set both "Reference alleles" and "Non-PASS variants" to "Dimmed" (figure 8) or "Hidden". These settings can be saved and used by default by any browsers.
- Configure the advanced filtering of the variant table view to contain two filters:
  - Reference allele = No
  - Filter = PASS

Optimize variant detection

Template workflows typically require customization to improve the performance of variant detection for a specific application. The optimal settings depend on factors such as the expected variant allele frequencies, the sequencing depth, and whether the variants of interest are somatic or germline.

Variant detection is performed in two steps:

- Variants are detected using an appropriate variant detection tool. The workflow used here is designed for detecting somatic variants and it therefore uses the Low Frequency Variant Detection tool.
- Detected variants are filtered using Filter on Custom Criteria to remove variants that are likely to be artifacts due to technical noise or sequencing errors.

An important concept when adjusting variant detection is the balance between recall, also known as sensitivity, and precision. Recall refers to the ability to detect variants that are truly present in the sample, which corresponds to detecting true positives (TPs) while minimizing false negatives (FNs; true variants that are not detected). Precision refers to the ability to remove variants that are artifacts due to technical noise and sequencing errors, which corresponds to detecting TPs while minimizing false positives (FPs; detected variants that are not truly present in the sample). Increasing recall makes it more likely to detect TPs, but usually also increases the number of FPs. Increasing precision reduces the number of FPs, but may also remove TPs and therefore increase the number of FNs. This trade off becomes particularly important for low frequency variants, where the signal from a true variant can be similar in magnitude to technical noise and sequencing errors. The recall/precision balance is adjusted by changing settings in both the variant detection tool and filter criteria.

When working with reference materials, variants present at a given frequency are expected to be detected at approximately the same frequency. However, due to the random sampling of DNA fragments during library preparation and sequencing, the observed frequency may vary. To

account for this variability, we recommend setting the **Minimum frequency (%)** lower than the target variant allele frequency. For example, when aiming to detect variants at 1%, the minimum frequency can be set to 0.5% to allow for expected fluctuations in read counts and coverage.

Assess the recall and precision

The tutorial data contains a workflow (figure 9) for benchmarking the variant detection results, i.e., assessing the recall and precision by comparing the detected variants to the verified reference material variants (table 1). The workflow uses Filter against Known Variants with different options and known variant tracks to determine the number of TPs, FPs, and FNs.

To refine the FPs, variants located outside the HG002 high confidence regions are removed using Filter Based on Overlap. In addition, Filter on Custom Criteria is used to restrict FPs to non-reference alleles that passed all filters. To refine the FNs, variants located outside the panel target regions are removed, ensuring that only variants expected to be detectable by the panel are included in the final FN set.

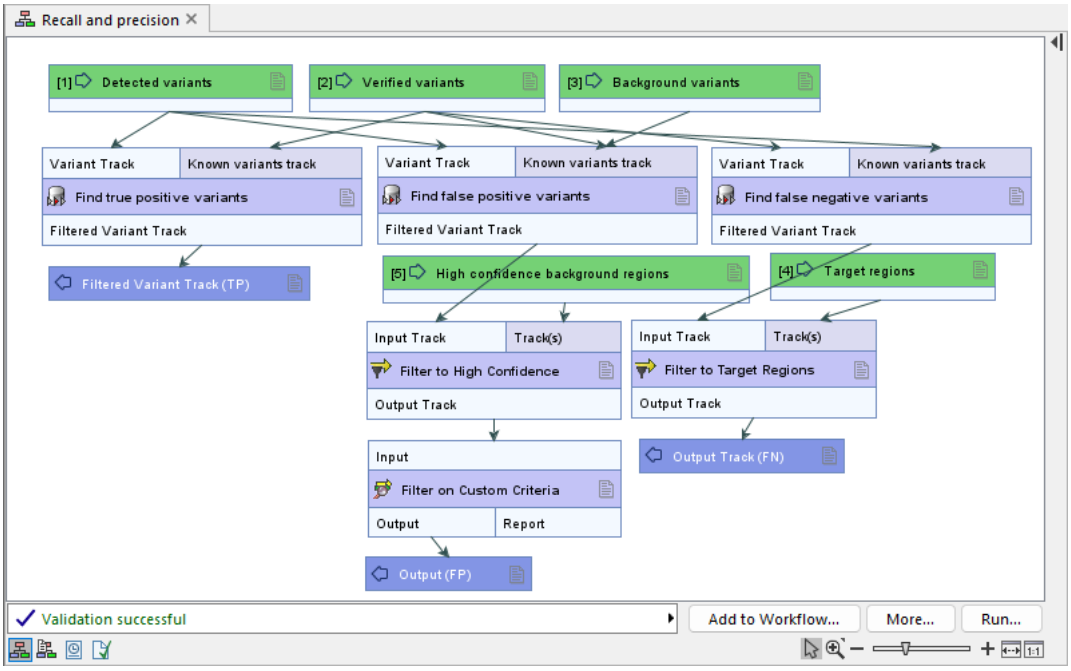

Figure 9: *Workflow for assessing the recall and precision.*

We will now run the workflow:

1. Open the imported workflow by double clicking on its name in the "Variant_detection" folder.
2. To launch the workflow, click on **Run...** located in the bottom right corner (figure 9).
3. In the first wizard step, **Select Detected variants**, select **Variants reads**, located in the "Results" folder.
4. In the **Select Verified variants** step, select **Seraseq_variants_Chr13+19**, located in the "Variant_detection" folder.
5. In the **Select Background variants** step, select **HG002_Chr13+19**, located in the "Variant_detection" folder.

6. In the **Select High confidence background regions** step, select **HG002_high_confidence_Chr13+19**, located in the "Variant_detection" folder.
7. In the **Specify reference data handling** step, select the "Variant Detection" Reference Data Set under "QIAGEN Tutorial".
8. In the **Target regions** step, keep the default settings.
9. In the last step, make a new subfolder in "Variant_detection" called "Recall and precision" and choose to save the workflow results there.

   Click on **Finish**.

When the workflow has finished executing, three variant tracks containing the true positive, false positive, and false negative variants (**TP**, **FP**, and **FN**, respectively) are located in the "Recall and precision" folder. We will now inspect these tracks to understand how to optimize the variant detection.

Open the **TP** track and click on the table view ( ) in the bottom left corner. All six expected variants are present in the track. However, one of them did not pass the filters (figure 10). Note that the TP track contains all detected variants that match the reference variants, regardless of whether they passed the variant filters. A reference variant that fails the filters is therefore classified here as a TP, but in practice represents a FN when assessing the recall.

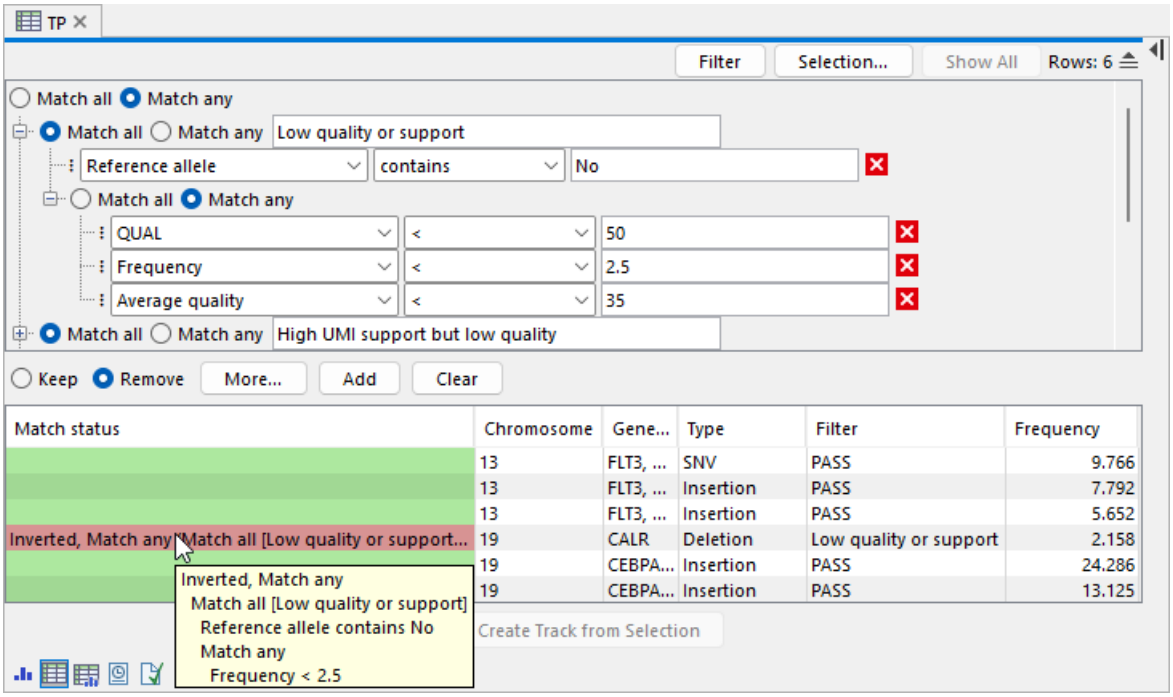

Figure 10: *The table view of the TPs. One variant did not pass the filters because of low quality or support. The "Filter variants" set is loaded and a detailed match status is displayed. The match status tooltip preserves the group structure of the filter set. The table view does not show all columns. The order of the visible columns has been changed by dragging the column names in the Side Panel. The Side Panel is collapsed.*

To better understand why this TP did not pass the filters:

- Apply the previously saved filter set (step 6):

1. Click on the downwards pointing arrow ( ) located in the top right corner of the filtering area to open advanced filtering.
2. Click on **More...**.
3. Click on **Filter Sets**.
4. Click on **Filter variants**.

- This immediately applies the filter to the table. To display again all variants, click on **Show All**.
- To display details on why the variant did not pass the filters, click on **More...** and **Add detailed match status**.

Hovering the mouse over a "Match status" cell reveals a tooltip with the exact criteria responsible for the variant not passing, while preserving the group structure of the filter set. It is now clear that the variant has a frequency that is too low to pass. We used a frequency of 2.5% when launching the workflow, which is half of the minimum frequency of the variants in the reference material (table 1). However, the actual frequency of this variant in the data is only 2.16%. This exemplifies the expected fluctuations in read counts and coverage and the imperfect relation between the expected and detected frequency, and illustrates why a minimum frequency threshold that is too strict can reduce recall even when all expected variants are present in the data.

We will now prepare the View Area for jointly examining the TP and FP variants.

Remove the match status by clicking on **More...** and **Remove match status**, and then close the advanced filtering by clicking on the upwards pointing arrow ( ) located in the top right corner of the filtering area.

Now open the **FP** track, switch to the table view, and drag it to a split view such that both TP and FP table views are visible at the same time (figure 11). To preserve TPs while removing FPs, we need to identify variant annotations that distinguish false positives from true positives. The following annotations, most of which are already used by the template workflow filtering step, are useful to examine: average quality, QUAL, coverage, count, frequency, read position test probability, read direction test probability, and proportion of singleton UMIs. For the tutorial data, the following can be observed:

- Average quality: most FPs have a lower average quality than the TPs.
- Read direction test probability: all FPs have a probability that is lower than that of the TPs.
- Proportion (singleton UMIs): most FPs have values close to 100%, while the TPs have lower values. This indicates that most FPs originate from UMI groups consisting of a single read, where sequencing errors are more likely to occur.

Finally, open the **FN** track. For the tutorial data, this track is empty, meaning that all expected variants within the panel target regions were detected by **Low Frequency Variant Detection**. This indicates that the current tool options achieve full recall, and it is only the filtering that needs updating. If the FN track were not empty, the options in the variant detection tool itself would need to be adjusted. Here, recall is limited only by the subsequent filtering.

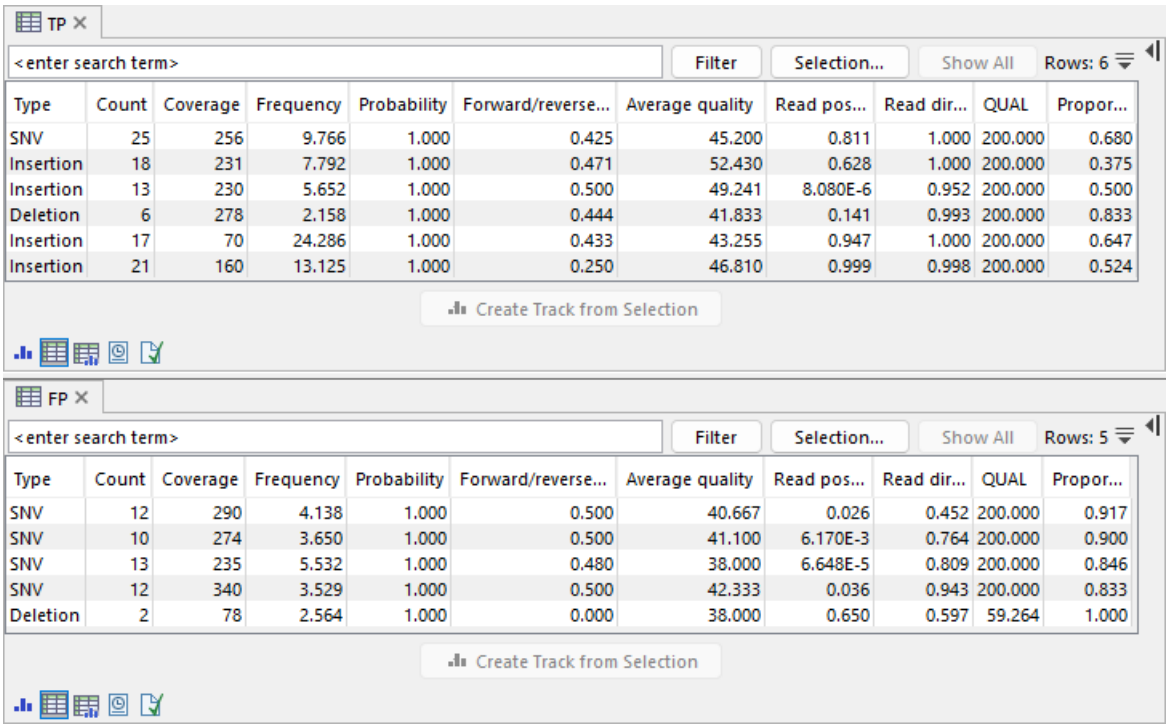

Figure 11: *Split view showing the table views of the TPs at the top and FPs at the bottom. The table views do not show all columns. The order of the visible columns has been changed by dragging the column names in the Side Panel. The Side Panel is collapsed.*

Update the variant filters

Based on the differences observed between TP and FP variants, we will now update the saved filter set to optimize the variant detection:

1. Open advanced filtering in the **TP** variant track by clicking on the downwards pointing arrow ( ) located in the top right corner of the filtering area.

   The "Filter variants" filter set is already loaded.

2. Update the frequency filtering to use 2 instead of 2.5.

3. Update the average quality filtering to use 40 instead of 35.

4. Collapse the "Low quality or support" group.

5. Add a new filter criterion by clicking on **Add**. The new criterion is added at the bottom, so scroll down so it is visible.

6. Configure the criterion to "Proportion (singleton UMIs) > 0.85".

7. Add a "Read direction test probability < 0.95" filter criterion.

8. Save the filter set by clicking on **More...**, **Filter Sets**, and **Save Filter...**, and use the previous name "Filter variants".

We will now apply the updated filter set to both the TP and FP variant tracks to verify that no true positives are removed and that all false positives are removed:

1. In the **TP** variant track, click on the **Filter** button at the top.
2. In the **FP** variant track, apply the saved filter set as previously described.

For the tutorial data, the read direction test probability alone would be sufficient to separate the FPs from the TPs. However, such clear separation should not be expected in general. When optimizing variant filters, it is important to avoid tailoring the filter criteria too specifically to a single dataset, as this may reduce performance when applied to other samples or sequencing runs. In practice, robust filter sets are typically based on a combination of multiple criteria and may require iterative adjustment using additional reference materials or representative datasets.

Update the workflow

The updated saved filter set can be used when running the template workflow in the **Filter on Custom Criteria** wizard step by clicking on **More...**, **Filter Sets**, and selecting the "Filter variants" filter set. If you run any template workflow on your own data, we recommend creating a copy of the template workflow and customizing it. See the Modifying a Workflow tutorial for examples of how to do so. In this case, the **Filter on Custom Criteria** workflow element should be updated to use the saved filter set.

Upload results to QCI Interpret

For clinical interpretation and reporting, the results can be uploaded to QCI Interpret. For this, you need an active API-enabled QCI Interpret or QCI Interpret Translational account.

The Biomedical Genomics Analysis plugin offers several QCI Interpret Upload tools. When launching the template workflow, in the **Specify workflow path** wizard step, set "Prepare for QCI Interpret" to "Yes". The workflow will produce a report containing all analyses performed that can be uploaded to QCI Interpret, including CNVs, TMB score, and MSI status, if these are chosen in the **Specify workflow path** step. This report can then manually be uploaded to QCI Interpret using Upload Prepared QCI Interpret Report.

Alternatively, the results can be uploaded using Upload to QCI Interpret. This can be useful if performing additional filtering of the variants produced by the workflow.